DOCUMENT RESUME

ED 295 971 TM 011 772

AUTHOR Micceri, Theodore
TITLE Estimating the Reliability of the CITAR Computer Courseware Evaluation System.
PUB DATE Sep 87
NOTE 57p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Computer Assisted Instruction; Computer Software; *Courseware; *Instructional Material Evaluation; Programed Instructional Materials; *Reliability
IDENTIFIERS *CITAR Computer Courseware Evaluation Model

ABSTRACT

In today's complex computer-based teaching (CBT)/computer-assisted instruction market, flashy presentations frequently prove the most important purchasing element, while instructional design and content are secondary to form. Courseware purchasers must base decisions upon either a vendor's presentation or some published evaluator rating. Unfortunately, these are almost certain to be biased by irrelevant factors such as color graphics or presentation speed. The Center for Interactive Technologies, Applications and Research (CITAR) Computer Courseware Evaluation Model (CCCEM) emphasizes the instructional components of such courseware, rather than the "bells and whistles" of the associated technology and provides descriptive information on more than 300 courseware components. Additionally, over 200 item-level tallies are synthesized into scores that may be used to compare similar packages on their instruction, management, physical, and presentational aspects. A study of the model's consistency found, on average, that seven of eight evaluators tend to agree on items and scores and that reliabilities for key scores are near 0.70. A comparison of traditional perceptual evaluations for the same courseware produced reliabilities averaging over 0.40 less than the CCCEM model, and agreements averaging 0.20 less. The results of this study suggest that the CCCEM fills the need of courseware purchasers for an objective, generic measure. Fifteen tables are included, and the distribution of CITAR scores attained by current CBT packages and item content of scales and sub-scales are appended. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295971

Estimating the Reliability of the CITAR Computer Courseware Evaluation System

by

Theodore Micceri

September, 1987

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

THEODORE MICCERI

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM011772

CONTENTS

Chapter	page
EXECUTIVE SUMMARY	1
BACKGROUND AND METHODS	4
Issues Arising in Scale Development and Interrater Agreement	8
Scaling	10
Scaling Items	14
Weighting	15
Interrater Agreement	18
Score Reliability	20
Sample	21
RESULTS	23
Item Level Agreement	23
Scale Level Percentages of Agreement	31
Reliability of Major Scores	35
ASSIGNMENT OF VALUES TO SCORES	37
Comparison of Rater Perceptions and CCCEM Scores	38
BIBLIOGRAPHY	40

Appendix	page
A. DISTIBUTION OF CITAR SCORES ATTAINED BY CURRENT CBT PACKAGES	41
B. ITEM CONTENT OF SCALES AND SUBSCALES	47

ABSTRACT

In today's complex CBT/CAI market, flashy presentations frequently prove the most important purchasing element while instructional design and content take a back seat to form. Courseware purchasers must base decisions upon either a vendor's presentation or some published evaluator rating. Unfortunately, these are almost certain to be biased by irrelevant factors such as color graphics or presentation speed. The CITAR courseware evalaution model (CCCEM) emphasizes the instructional components of such courseware rather than the "bells & whistles" of the associated technology and provides descriptive information on more than 300 courseware components. Additionally, over 200 item-level tallies are synthesized into scores that may be used to compare similar packages on their Instruction, Management, Physical and Presentation aspects. A study of the model's consistency found on average that 7 of 8 evaluators tend to agree on items and scores, and that ICC and test-retest reliabilities for key scores were near .70. A comparison of traditional perceptual evaluations for the same courseware produced reliabilities averaging over .40 less than the CCCEM model, and agreements averaging about .20 less. The results of this study suggest that the CCCEM fills the need of courseware purchasers for an objective, generic measure.

EXECUTIVE SUMMARY

Funded by a grant from the Westinghouse Corporation, the Center for Interactive Technologies, Applications and Research (CITAR) at the University of South Florida initiated in July, 1986, an attempt to develop an objective, generic model for the evalaution of computer based courseware. The CITAR Computer Courseware Evaluation Model (CCCEM) attempts to break conceptually complex concepts such as instruction, management and user interface into component pieces small enough to be objectively defined. In this way, both the existence and prevalence of various content and technological aspects of computer courseware may be compared across products.

The CCCEM contains information on over 250 facets of CBT package courseware. Faced with this plethora of information the harried consumer will most likely base his/er purchasing decision almost entirely on an unreliable evaluator rating. Unfortunately, synthesizing so many diverse phenomena is quite difficult, therefore overall ratings based soley on evaluator interpretations more likely indicate personal biases than TRUE differences in software effectiveness. CCCEM offers an alternative in the form of absolute scores on which the relative efficiency of different educational soft-

ware may be compared. The items included in these scores may be conceived of as stylistic or generic in the sense that their presence should associate with improved learning for almost any subject area. Further, the presence of more generic items should associate with "greater" learning than the presence of fewer such items.

Average agreement among raters across the several major areas of content (Physical, Presentation, Management and Instruction) produced interrater agreement percentages showing that approximately nine of ten raters code the same (above 80% agreement). Thus, overall agreement appears quite adequate. Certain specific items (i.e. in Questioning, Sequencing and Response Judging) produced lower percentages, showing agreement among as few as three of four raters.

Estimates of reliability (an instrument's ability to consistently discriminate among CBT packages) were adequate for Efficacy, Management and Instruction scores (circa .70), and acceptable for Physical scores (.55). However, the Presentation score is not adequately reliable for decision purposes (.15). The modest reliabilities exhibited by the Presentation and Physical scores result largely from the limited variability found among CBT packages in these areas. AS software becomes more sophisticated and efficacious, one would expect variability and therefore reliability to increase. Given the current levels of variability among CBT

packages, it appears appropriate to separate the **Outstanding** from the **Excellent**, the **Moderate**, and the **Questionable**, on Efficacy, Instruction, Management and Physical characteristics.

Additional study, directly comparing simple rater perceptions with the relatively objective CCCEM scores suggests that the CCCEM exhibits a substantial advantage both in agreement and reliability. In fact, CCCEM reliabilities averaged .47 higher than rater perceptions for the same questions.

BACKGROUND AND METHODS

This document presents the results of a study conducted to determine the operational consistency (reliability) of the Center for Interactive Technologies, Applications and Research (CITAR) Computer Courseware Evaluation Model (CCCEM). By definition, a reliable instrument exhibits stability, consistency, dependability and small errors of measurement on the characteristic begin measured. Although all measures include some measurement error, judgements made by humans are especially plagued by this problem. Thus, clearly delineated judgement criteria become a primary concern of instrumentation involving human decisions. This study examines the consistency with which judgements on 10 pieces of educational software are made by a set of four raters using CCCEM.

Reliability may be viewed as the consistency with which an instrument differentiates among a group of targets (CBT packages). Put simply, the consistency with which the same relative rankings are assigned to a specific group of targets by different raters. A major source of error for observation instruments is disagreement among raters, thus this issue was considered separately from the overall reliability question to isolate its influence. Interrater

agreement may be defined as the extent to which two or more observers, working independently, agree on which phenomena occur to what degree in the target of interest.

Mitchell (1979) claims the most common index of rater consistency in observational studies to be the interobserver agreement percentage. This technique, although an excellent measure of the absolute magnitude of one type of error (observer disagreement), provides no information about an instrument's ability to differentiate among targets. therefore, extremely high interrater agreement percentages may associate with very low reliabilities and vice versa. Although oft criticized as a measure of consistency (see Towstopiat, 1984; Frick & Semmel, 1978), mean item level percentage of agreement provides an accurate and readily interpretable measure of rater consistency for the data produced by the CCCEM model, particularly at the item level. Therefore, mean percentages of agreement were used as an estimate of rater consistency.

Although numerous techniques have been suggested for the investigation of observation instrument reliability, consensus appears to support the use of generalizability theory and the intraclass correlation coefficient (ICC) based on Analysis of variance (Mitchell, 1979; Shrout and Fleiss, 1978; Tinsley and Weiss, 1975). Tinsley and Weiss (1975) support the use of the intraclass correlation (ICC) as an

estimate of rating scales reliability because it permits the inclusion or exclusion of the between-rater variance as part of the error variance (WMS) and because it allows an estimation of the precision of the reliability coefficient (p. 363).

Since every statistical estimate contains unique biases (Cook and Campbell, 1984), a second, equally conservative estimate of reliability was also computed: average test/retest among raters. By taking the average correlation of scores obtained among CBT packages for each rater with those of every other rater, this estimate indicates the degree to which different raters rank the same CBT packages in the same order, and as with the ICC, includes almost every source of variance in the error term.

To estimate the reliability of CCCEM, one must asses its ability to consistently discriminate among various CBT packages differing on items of interest. It should also fail to differentiate among those CBT packages not differing of items of interest. The issue therefore becomes: Which items are of interest?

Assuming at least two major market segments for CBT packages (managers/teachers and students), different aspects of these subscores interest different audiences. For example, management systems appeal more to teachers/managers than to students, while presentation and instructional aspects ap-

peal to both audiences. In addition, at least two major marketing considerations are (1) Instructional effectiveness and (2) Cost/Appeal/Physical characteristics. For example, although not necessarily more instructionally effective for highly motivated students, "Bells and Whistles" such as color graphics and animation may stimulate the interest of reticent learners and perhaps even appeal to the jaded tastes of experienced teachers/managers.

Five major scores were created from CCCEM items to service the needs of varying audiences:

1. Descriptive (not scalable)
   a) All specific characteristics of CBT package's,
2. Instruction
   a) Characteristics of the learning environment, for instance, questioning, problem solving or gaming,
3. Management
   a) Characteristics of the management system such as group level analysis and student performance reports.
4. Presentation
   a) Characteristics of physical interface between student and CBT package.
5. Physical
   a) Characteristics of hardware and software.

Each subscore relates the particular CBT package against a hypothetical optimum CBT package regarding both software sophistication and instructional effectiveness.

ISSUES ARISING IN SCALE DEVELOPMENT AND INTERRATER AGREEMENT

Instrument validation requires both that items vary across CBT packages, and that this variability be related to greater and lesser instructional effectiveness. Two CTBs may differ in one area (e.g. hardware or management systems) but not in others (e.g instruction techniques). In addition, various "differences" within each cell of the CCCEM model will impact instructional effectiveness and target audience appeal in varying ways. Failure to weigh such considerations could substantially reduce both the usefulness and marketability of CCCEM's evaluations.

Scalable items within the CCCEM model are described in one of three ways. First is a binary Yes/No, defining existence or non-existence. Secondly, a scale including four points is used to relate the necessity of a particular characteristic to the use of a package. This scale is designated RONSI with points representing respectively: (R)equired, (O)ptional, (N)ot (S)upported or (I)ndeterminate. Most RONSI data are collected directly from vendor supplied materials. Thirdly, a four point Likert-type scale describes the quantity of or extent to which a characteristic appears in the CBT package. The items are designated ESON and rep-

resent respectively: (E)xtensively, (S)ignificantly, (O)ccasionally or (N)egligible.

The validity of an instrument depends upon relatively optimal weightings of item and subscores in the creation of composite scores. For instance, if one item (i.e. defined objectives) is scored on a yes/no basis and assigned values of 0 for no and 1 for yes, while another (i.e. use of several fonts) is scored using ESON and assigned values of 0 for Negligible and 3 for Extensively, it is obvious that more than one font would influence score variability to a greater degree than the presence or absence of defined objectives. This, although the latter is far more important both to instruction design and implementation. Thus, the weights assigned to each item and subscore were estimated by a panel of "experts". These weightings may later be validated by studies of CBT package effectiveness.

Secondly, the assignment of values to responses for RONSI and ESON may differ from item to item.

	E	S	O	N
Page Turning	1	1	2	0
Tutorials	2	2	1	0

The preceding assigns different values to different frequencies of occurrence for two ESON scaled items. Thus, the Extensive or Significant use of tutorials is given twice the value of its Occasional occurrence, while the Extensive or Significant use of page turning receives only half the value of its Occasional occurrence. Although these issues contain statistical aspects, they must largely be solved by expert opinion.

SCALING

Maximum scores represent what to the best current knowledge are relatively **OPTIMAL** CBT package learning environments. Four of the five previously mentioned subscores are scalable:

1. **Instruction** - consisting of 115 scalable items, 6 major subscales and 16 minor subscales,
2. **Management** - consisting of 51 scalable items and 9 subscales,
3. **Presentation** - consisting of 31 scalable items and 6 subscales and
4. **Physical** - consisting of 19 scalable items and 3 subscales.

These scales are amenable to combination into a total score which may be called: **Efficacy.**

During May and June, 1987, five meetings were held with "experts" representing the following interest groups: man-

agement, instruction design, software evaluators and measurement. Each meeting included at least five such experts. During this time, item scalings were tentatively determined for each of the CCCEM items. In addition, subscore composition, levels of weightings (item, minor subscale, major scale) were tentatively defined. Tables 1 and 2 list subscales, weighting and scoring characteristics of the four major submeasures, respectively: Instruction, Management, Presentation and Physical.

TABLE 1

Components of Instruction Scores

Major and Minor Subscales	# Items	Max Wt/Scr	Mean item Weight
PLANNING*			
1. Student Document	5	4	0.80
2. Task Analysis	1	6	6.00
	6	10	1.67
OBJECTIVES			
1. Defined Goals	2	30	15.00
2. Defined Outcomes	6	40	6.67
3. Defined Processes	6	5	0.83
sub totals	20	75	3.75
TECHNIQUES			
1. Instruction Techniques	10	60	6.00
2. Physical Present	4	35	8.75
3. Timing	5	5	1.00
sub totals	19	100	5.26
INTERACTION			
1. Judging incorrect responses	10	35	3.50
2. Testing	3	25	8.25
3. Interaction	16	5	0.31
4. Questioning	11	25	2.27
5. Sequencing	14	20	1.43
sub totals	54	110	2.04
RESOURCE SCOPE			
1. Intrinsic	5	50	10.00
2. Extrinsic	8	20	2.50
3. Supplemental	8	10	1.25
sub totals	21	80	3.81
CONTENT ASSESSMENT	**1**	**25**	**25.00**
Totals	115	385	3.34

* Items comprising each scale and subscale are detailed in Appendix B.

TABLE 2

Components of Management, Presentation and Physical Scores

Minor Subscales	# Items	Max Wt/Scr	Mean Item Weights
MANAGEMENT			
1. Documentation	8	10	1.25
2. Record Keeping	6	25	3.57
3. Data Analysis/ Entry	7	20	2.86
4. Reports	3	15	5.00
5. Media	4	15	3.75
6. Access	5	10	2.00
7. Security	2	10	5.00
8. Learning Prescriptions	5	10	2.00
9. Instructor Browse/ Sequencing	11	10	0.91
Totals	51	125	2.45
PRESENTATION			
1. Interface	6	1	0.15
2. Text	7	10	1.43
3. Graphics	7	10	1.43
4. Sounds	4	10	2.50
5. Design Features	4	10	2.50
6. Text editing	3	1	0.33
Totals	31	42	1.35
PHYSICAL			
1. Screens	7	30	4.29
2. Peripherals	11	10	0.91
3. Concurrent applications	1	10	10.00
Totals	19	50	2.63

Note: Items comprising each scale and subscale are detailed in Appendix B.

SCALING ITEMS

Within the subscales shown in Tables 1 and 2, each of the 222 scalable CCCEM items was evaluated separately to determine both its importance against all other items to which it relates, and an optimum application for each set of items (subscale). Every item receives a score from its rater with the following ranges:

Yes	No	R O NS I	E S O N
1	0	1 2 3 4	1 2 3 4

For most items, the lack of that item contributes nothing to either the management or instruction functions of the CBT package (No in the Yes/No dichotomy, Not Supported (NS) or Indeterminate under RONSI and Negligible for ESON. Thus, these were generally assigned a value of zero (0). For most items, Yes in the Yes/No dichotomy, Optional under RONSI, and either Occasional or Significant in ESON represents the optimal learning environment. For management subscales, Extensive (under Eson) was frequently considered optimal. These scores generally received high values. Required (RONSI) or Extensive (ESON) use of any technique may be considered sub-optimal in the learning environment, therefore, these usually received an intermediate value.

WEIGHTING

Since the CCCEM contains numerous individual items, and since the number of items in a possible area of Instruction, Management or Physical/Presentation may or may not relate to the importance of that area in an overall perspective, each of the noted areas were assigned relative weights in the production of a total **Efficacy** score: Instruction (64%), Management (21%), Presentation (7%) and Physical (8%). Within each of the major scores produced by CCCEM, several subscales exist, containing different numbers of items. The Instruction score contains six major subscales (Planning, Objectives, Techniques, Interaction, Resource Scope and Content Assessment) and 17 minor subscales. Each of the major subscales was weighted relative to the other five as follows: Planning (2.5%), Objectives (19.5%), Techniques (26%), Interaction (28.6%), Resource Scope (20.8%) and Content Assessment (6.5%). Planning and Content Assessment were assigned low weights due to their current incomplete status.

Within each of the major subscales, its minor subscales was assigned a weight relative to each other minor subscale as shown in Tables 1 and 2.

Within each of the minor subscales, each of the items comprising it was assigned a weight relative to each other item (see Appendix B for detailed item definition). For ex-

ample, the six items in the minor subscale **Defined Outcomes** were assigned the following values: Cognitive (2), Measured (2), Affective (1), Measured (2), Psychomotor (2), Measured (2).

In creating scores and subscores, the relative weightings assigned to each item were multipled times the relative values assigned to that item. The sum of values created by this process for each minor subscale was then computed and divided by the maximum possible score for the minor subscale (defined as an optimal situation - this need not include all items) and multiplied times the weighting assigned to that minor subscale. The sum of minor subscale values created in this fashion was used to compute major subscale values, and the sum of major subscale values taken as the value for that score.

For example, the 10 items of the Instruction Techniques minor subscale of the Techniques major subscale of the Instruction score is composed of ESON items which are scaled and weighted as follows:

Item	Scaling				Weighting
	E	S	O	N	
Page Turning	0	1	2	0	0
Tutorial	1	2	2	0	3
Drill & Practice	1	2	2	0	2
Questioning	1	2	2	0	1
Concurrent Training	2	2	1	0	4
Problem Solving	1	2	2	0	3
Simulation	2	2	1	0	4
Modeling	1	2	2	0	2
Gaming	1	1	2	0	2
Inquiry	2	2	1	0	3

The assumed optimum situation for Learning for this subscale would include the Occasional or Significant use of perhaps three or more higher weighted subscales, in addition to expected items such as page turning or questioning. An "optimal" environment would score approximately 12 points. Thus, the total score obtained by each CBT package would be divided by 12 to determine the proportion of an optimal environment contained within the CBT package. For example, a CBT package using Significant Page Turning (.5 X 0 = 0.0) plus Occasional Drill & Practice (1.0 X 2 = 2) plus Significant Problem Solving (1.0 X 3 = 3) plus Occasional Simulation (0.5 X 4 = 2) would sum to 0.0 + 2.0 + 3.0 + 2.0 = 7.0 and would receive a score of 7.0 /12 = .583, or 58% of the optimal environment. This percentage would then be multiplied times the maximum subscale value for Instruction Technqiues (.58 X 60 = 35) and this value (35) would be summed with the

CBT package's score for Physical Presentation and Timing to create an Instrucational Techniques major subscale score (Table 1). This score would then be combined with the other major subscales under Instruction to derive a total Instruction score. In addition, the proportion of "optimality" for a specific subscale, in this case .58, allows for absolute comparisons on a scale from zero to 100, where zero represents "Useless" and 100, "Optimum".

Interrater Agreement

Mean percentages of agreement were computed for each of four raters for each of 10 tapes. These means were then considered both independently and in the form of sub and total scores within instrument domains. When computing the arithmetic mean for percentages (or any other ratio), a value of .50 indicates the numerator was twice as small as the denominator, a value of .25 four times as small and a value of .10, ten times as small. If the percentages are simply averaged, the mean would be .283, and would suggest agreement about one fourth the time. This is incorrect, since the actual multiples are respectively 2, 4 and 10 (mean = 5.33). Thus, the mean percentage should be .188 or agreement about one fifth, not one fourth of the time. To avoid this problem, the log to the base e of each individual percentage was used in the computation of mean logs. The anti-log of this mean was then taken to define the percentage of agreement for the "average" rater.

Traditionally, agreement percentages are computed at the item level, assigning an agreement percentage of 100 for each item of agreement and of 0.00 for each disagreement and computing the mean. Although appropriate for yes/no and RONSI items, this technique must underestimate interrater agreement for the S and O elements of the ESON scale. Obviously, E(xtensively) agrees with S(ignificantly) to a greater degree than it does with N(egligible). Therefore, for all ESON scaled items, adjacent pairs (E and S, S and O, O and N) were considered to agree 50%. All other disagreements were assigned a value of 0% agreement. Such a scaling does not specify a model of instruction, but rather credits various possible combinations

Since both item level (for descriptive purposes) and score level (for purposes of ranking CBT packages) are of interest here, agreement was computed at both levels. At the item level, percentage of agreement represents the mean agreement among four raters (six rater pairs). For composite scores, agreement represents the mean percentage across all items (i), all rater pairs (k_r) and all subjects (j) for each score, with item level agreement defined as:

$$1.0 \qquad R = \frac{x_i - x_{(i+1)}}{\text{Range}}$$

SCORE RELIABILITY

For the CCCEM, the average reliability of an individual judge is of interest, since each CBT package will be rated by only one judge. Thus, mean differences in the ratings of different judges is of great importance. Also, one wishes the results of these analyses to be generalizable to other samples of raters using the same scale with a similar sample of targets. Therefore, between rater variance is here included in the error term, which gives the expected reliability of an "average" judge and reduces the reliability estimate toward its lower bound (Shrout & Fleiss, 1978). The ICC coefficient is computed using an Analysis of variance table: with between target variance (BMS) treated as the "True variability" and within cell variance (WMS - combining instrument, time and rater error) treated as error:

$$2.0 \qquad R = \frac{BMS - WMS}{BMS + (k-1)WMS}$$

where:

BMS = between targets mean square
WMS = error or within targets mean square
k = number of raters/judges

A second, equally conservative reliability estimate was produced by taking the median Pearson Product Moment Correlation Coefficient among each pair of raters scores for all

10 CBT packages. This estimate contains different sources of error than the preceding, and provides a good indication of the consistency with which different raters rank the same CBT packages in the same order from high to low.

SAMPLE

Educational software at present provides only limited examples in the academic disciplines. Much software is dedicated to various word processing and database management systems in addition to some traditional areas of skill development such as typing. Therefore, a sample was chosen to represent certain strata (characteristics) present in the population of all CBT package courseware. Table 3 shows the selected sample An attempt was made to include several commonly occuring subject areas, several major producers of software, several contact time intervals and several levels of instruction design complexity. At the time of sampling, neither the range nor variability of CCCEM scores was known. Four experienced raters evaluated each CBT package.

TABLE 3

Sample of Software Chosen as Targets for Reliability Study

Publisher	Title/Subject	Contact Time	Efficacy Score
Blue Chip	American Dream/Management	8	55
ATI	DBASE III	2	44
ATI	SUPERCALC	2	35
Compre	Reading and Understanding	20	36
CDEX	IBM PC DOS	4	47
CDEX	LOTUS III	2	49
David	Sharpening Your Executive Writing Skills	3	54
Thoughtware	Managing Stress	4	46
Thoughtware	Improving Employee Performance	4	50
QED	Typing Made Easy	4	58

RESULTS

ITEM LEVEL AGREEMENT

The following discussion considers only the 192 items upon which disagreement could occur. The remaining items (24 scalable) derive from CBT package documentation and are not subject to rater disagreement.

Within these data, one major source of error (disagreement) exists that usually is not present during evaluations. To obtain an adequate number of experienced evaluators for the reliability study, it proved necessary to include evaluators having different areas of expertise (2 MBAs, 1 Psychologist, 1 Linguist). Thus, for every CBT package evaluated, at least one evaluator was out of field. This is particularly important when one considers that agreement among three of four evaluators produces a percentage of 50% (not 75%). This results because six possible agreement pairs occur for four evaluators (1 & 2, 1 & 3, 1 & 4, 2 & 3, 2 & 4 and 3 & 4). Thus, if one evaluator codes R(onsi) while three code O(ptional), only three of the six pairs agree (50%) although 75% of the evaluators agree that O(ptional) is correct. Assigning partial agreement (50%) for adjacent ESON scale points reduces this problem some-

what, but for all agreements, the percentage obtained is, both conceptually and realistically, something of an underestimate. Table 4 shows how many raters must agree to produce various percentages of agreement. It is clear from Table 4 that 75% agreement is quite good. Thus, in the following discussions, 75% or greater agreement will be considered adequate. The assignment of partial agreements for ESON scalings increases the percentage of agreement somewhat (for four raters, the increase averages approximately 10% to 15%, since many disagreements are not adjacent scale points).

TABLE 4

Interpreting Agreement Percentages

Number of Evaluators Reaching Agreement	Percentage of Agreement
3 of 4	50%
4 of 5	60%
5 of 6	67%
6 of 7	71%
7 of 8	75%

Another, unanticipated source of error resulted from a definition clarification session which took place during the conduct of reliability evaluations by CBT package evaluators. Definitions were clarified on several heavily weighted items. Later Discussions with specific evaluators indi-

cated that, following the session, their codes would more closely corresponded with other evaluators for certain items where differences exist in the reliability data reported here. It was deemed inappropriate, however, to make any adjustments in the data for these situations. Therefore, the reliabilities reported are conservative for yet another reason.

Table 5 shows the item level agreement percentages for the 192 scalable items upon which disagreement may exist. In general, agreements are high, with only 20.8% showing agreement among fewer than seven of eight evaluators. Some areas of relatively great disagreement may be attributed to evaluators working in areas outside their expertise. In particular, whether books are an Extrinsic (67%) or a Supplemental (55%) aspect of the resource scope; the frequency of Context Sensitive Questions (72%); and the techniques used for Response Judging (62%): Only One Response (57%), Wait for Correct Response (63%), Wrong and Continue (69%), Wrong and Hint (72%), Give Correct and Go (56%).

Some items appear to require, and since have undergone clarification or simplification: (1) Documentation - Instructor (50%), Student (50%) and whether it is in the form of Manuals, Tutorials or Both (45%), (2) Varying - Text Fonts (73%) and Formats (52%), (3) Questioning - Continuation Response (69%) and On-Line Help (65%), (4) Graphics -

whether Color (58%), Windowing (66%) or Illustration (66%), and, (5) Sequencing, where several items fall below 75% agreement.

At least one item appeared difficult for evaluators to determine: Security - Student Access (70%). Perceptions regarding the preponderence of certain items also appear to differ somewhat, suggesting the need for clearer definitions: Page Turning (72%) and Questioning (69%) as instruction techniques; Forced Waits (70%) and ability of students to Quit (60%) lessons at any time; and Feedback (50%): Positive (62%), Negative (75%), Neutral (58%). In addition, one item showed a relatively low percentage of agreement for no apparent reason: preponderance of Computer Tones (53%).

TABLE 5

Item Level Agreement for Four Raters Across Ten CBTs

Item	CBTs n	Ratrs n	Mean Pct. Agree
COLOR	10	4	1.00
80 CHARACTERS	10	4	0.90
40 CHARACTERS	10	4	0.93
UPPER AND LOWER CASE	10	4	1.00
PERIPHERALS	10	4	0.93
PERIPHERALS PRINTER	10	4	0.90
DOCUMENTATION MANAGER	10	4	0.50
DOCUMENTATION MANUALS	10	4	0.83
DOCUMENTATION TABLES OF CONTENTS	10	4	0.80
DOCUMENTATION INDEX	10	4	1.00
DOCUMENTATION INSTRUCTOR	10	4	0.50
DOCUMENTATION MANUALS	10	4	0.88
DOCUMENTATION TABLES OF CONTENTS	10	4	1.00
DOCUMENTATION INDEX	10	4	1.00
DOCUMENTATION STUDENT	10	4	0.50
DOCUMENTATION MANUALS	10	4	0.45
DOCUMENTATION TABLES OF CONTENTS	10	4	0.93
DOCUMENTATION INDEX	10	4	1.00
MS RECORD KEEPING?	10	4	0.80
INDIVIDUALS	10	4	0.84
GROUPS	10	4	0.92
SUBGROUPS	10	4	1.00
TESTING HISTORY	10	4	0.88
UNANTICIPATED RESPONSES	10	4	0.90
COURSEWARE VALIDITY	10	4	1.00
MS DATA ENTRY AUTOMATIC	10	4	0.80
MS DATA ENTRY MANUAL	10	4	1.00
MS DATA ANALYSIS	10	4	0.95
MS DATA ANALYSIS INTERNAL	10	4	0.95
MS DATA ANALYSIS TEST	10	4	0.95
MS DATA ANALYSIS ITEM	10	4	0.95
MS DATA ANALYSIS EXTERNAL	10	4	0.95
MS REPORTS	10	4	0.83
MS REPORTS PRE FORMATTED	10	4	0.83
MS REPORTS USER FORMATTED	10	4	1.00
MS MEDIA PAPER BASED	10	4	0.92
MS MEDIA FLOPPY	10	4	0.72
MS MEDIA NETWORK	10	4	1.00
MS MEDIA MAINFRAME	10	4	1.00
MS ACCCEMSS	10	4	0.75
MS ACCCEMSS TERMINAL	10	4	0.95
MS ACCCEMSS MICRO	10	4	0.75
MS ACCCEMSS HARDCOPY	10	4	0.76

Table 5 Continued
Item Level Agreement for Four Raters Across Ten CBTs

Item	CBTs n	Ratrs n	Mean Pct. Agree
MS SECURITY	10	4	0.90
MS SECURITY STUDENT ACCESS	10	4	0.70
MS ACCCEMSS MANAGER ONLY	10	4	0.75
MS LEARNING PRESCRIPTIONS	10	4	0.93
MS LEARNING PRESCRIPTIONS FORMATIVE	10	4	0.95
MS LEARNING PRESCRIPTIONS COURSE LEVEL	10	4	0.95
MS LEARNING PRESCRIPTIONS MODULE LEVEL	10	4	1.00
MS LEARNING PRESCRIPTIONS ADAPTATION	10	4	0.95
OUTCOME DEFS COGNITIVE	10	4	1.00
OUTCOME DEFS MEASURED	10	4	0.93
OUTCOME DEFS AFFECTIVE	10	4	1.00
OUTCOME DEFS MEASURED	10	4	1.00
OUTCOME DEFS PSYCHOMOTOR	10	4	1.00
OUTCOME DEFS MEASURED	10	4	1.00
PROCESS DEFS COGNITIVE	10	4	1.00
PROCESS DEFS DOCUMENTED	10	4	1.00
PROCESS DEFS AFFECTIVE	10	4	1.00
PROCESS DEFS DOCUMENTED	10	4	1.00
PROCESS DEFS PSYCHOMOTOR	10	4	1.00
PROCESS DEFS DOCUMENTED	10	4	1.00
RESOURCE SCOPE INTRINSIC	10	4	0.78
TEXT	10	4	0.75
GRAPHICS	10	4	0.84
ANIMATION	10	4	0.88
SOUND	10	4	0.77
SOUND ON/OFF	10	4	0.83
RESOURCE SCOPE EXTRINSIC	10	4	0.77
BOOKS	10	4	0.67
LABS	10	4	1.00
SEMINARS	10	4	1.00
LECTURES	10	4	1.00
FIELD TRIPS	10	4	1.00
HELP SESSIONS	10	4	1.00
TECHNOLOGICAL	10	4	0.95
SUPPLEMENTAL	10	4	0.77
BOOKS	10	4	0.55
LABS	10	4	1.00
SEMINARS	10	4	1.00
LECTURES	10	4	1.00
FIELD TRIPS	10	4	1.00
HELP SESSIONS	10	4	1.00
TECHNOLOGICAL	10	4	0.76

Table 5 Continued
Item Level Agreement for Four Raters Across Ten CBTs

Item	CBTs n	Ratrs n	Mean Pct. Agree
PAGE TURNING	10	4	0.72
TUTORIAL	10	4	0.81
DRILL AND PRACTICE	10	4	0.90
QUESTIONING	10	4	0.69
CONCURRENT TRAINING	10	4	0.91
PROBLEM SOLVING	10	4	0.92
SIMULATION	10	4	0.77
MODELING	10	4	0.87
GAMING	10	4	0.90
COOPERATIVE GAMING	10	4	1.00
INQUIRY	10	4	0.91
TEXT COLORS	10	4	0.76
TEXT SIZING	10	4	0.81
TEXT FONTS	10	4	0.73
TEXT FORMATS	10	4	0.52
TEXT FLASH OR BLINK	10	4	0.87
TEXT PRINTOUTS	10	4	0.83
TEXT SCREEN DUMP	10	4	0.83
GRAPHICS COLOR	10	4	0.58
GRAPHICS ANIMATION	10	4	0.84
GRAPHICS VIDEO	10	4	1.00
GRAPHICS WINDOWING	10	4	0.66
GRAPHS, CHARTS ETC	10	4	0.86
ILLUSTRATIONS	10	4	0.66
GRAPHICS FLASH OR BLINK	10	4	0.94
GRAPHICS PRINTOUT	10	4	0.95
GRAPHICS SCREEN DUMP	10	4	0.95
SOUNDS COMPUTER TONES	10	4	0.53
SOUNDS EXTERNAL SOURCES	10	4	0.92
SOUNDS MUSIC	10	4	0.92
SOUNDS SYNTHESIZED VOICE	10	4	1.00
TIMING FORCED WAIT	10	4	0.70
OVERRIDE	10	4	0.95
FORCED MOVEMENT	10	4	0.86
OVERRIDE	10	4	0.97
DISPLAY INTERRUPT	10	4	0.95

Table 5 Continued
Item Level Agreement for Four Raters Across Ten CBTs

Item	CBTs n	Ratrs n	Mean Pct. Agree
INTERACTION PROMPTS	10	4	0.77
INTERACTION FEEDBACK	10	4	0.50
INTERACTION FEEDBACK POSITIVE	10	4	0.62
INTERACTION FEEDBACK NEGATIVE	10	4	0.75
INTERACTION FEEDBACK NEUTRAL	10	4	0.58
INPUT QUITTING	10	4	0.60
INPUT RESTART	10	4	0.78
QUESTIONS?	10	4	0.78
QUESTIONS MULTIPLE CHOICE	10	4	0.78
QUESTIONS MATCHING	10	4	0.87
QUESTIONS TRUE FALSE	10	4	0.80
QUESTIONS FILL BLANK	10	4	0.75
QUESTIONS NUMERIC RESPONSE	10	4	0.82
QUESTIONS GRAPHIC	10	4	1.00
QUESTIONS FREEFORM TEXT	10	4	0.84
QUESTIONS CONTINUATION RESPONSE	10	4	0.69
QUESTIONS ON LINE HELP	10	4	0.65
QUESTIONS CONTEXT SENSITIVE	10	4	0.72
RESPONSE JUDGING	10	4	0.62
ONLY ONE RESPONSE	10	4	0.57
SLIGHT ERRORS	10	4	0.81
SENTENCE PARSING	10	4	1.00
LINGUISTIC	10	4	0.92
MATHEMATICAL	10	4	0.95
NATURAL LANGUAGE	10	4	1.00
WAIT FOR CORRECT RESPONSE	10	4	0.63
WRONG AND CONTINUE	10	4	0.64
WRONG AND HINT	10	4	0.72
WRONG AND HELP	10	4	0.96
WRONG AND TUTORIAL	10	4	0.97
WRONG AND OFF LINE	10	4	1.00
SCORE AND GO	10	4	0.84
GIVE CORRECT AND GO	10	4	0.56
STUDENT TYPES CORRECT RESPONSE	10	4	0.72
ANSWER KEY PROVIDED	10	4	0.82

Table 5 Continued
Item Level Agreement for Four Raters Across Ten CBTs

Item	CBTs n	Ratrs n	Mean Pct. Agree
SEUQENCE LINEAR	10	4	0.62
SEUQENCE DUMMY KEY	10	4	0.57
SEUQENCE BRANCHING	10	4	0.56
SEUQENCE DIRECTION	10	4	0.53
SEUQENCE BACKWARD?	10	4	0.60
SEUQENCE ONE PAGE BACKWARD	10	4	0.63
SEUQENCE MULTIPLE PAGES BACKWARD	10	4	0.92
SEUQENCE FOWARD	10	4	0.73
SEUQENCE ONE PAGE FORWARD	10	4	0.80
SEUQENCE MULTIPLE PAGES FOWARD	10	4	0.95
SEUQENCE MASTERY	10	4	0.92
SEUQENCE CREDIT FOR MASTERY	10	4	0.94
SEUQENCE ADAPTIVE	10	4	0.88
SEUQENCE BROWSE	10	4	0.70
SUPPLEMENTAL	10	4	0.77
BOOKS	10	4	0.55
LABS	10	4	1.00
SEMINARS	10	4	1.00
LECTURES	10	4	1.00
FIELD TRIPS	10	4	1.00
HELP SESSIONS	10	4	1.00
TECHNOLOGICAL	10	4	0.76
DESIGN FEATURES SCREEN LAYOUT	10	4	0.95
DESIGN FEATURES LESSON FLOW	10	4	1.00
DESIGN FEATURES COLOR	10	4	0.80
DESIGN FEATURES SOUND	10	4	0.75
TESTING RANDOM	10	4	0.84
TESTING FIXED	10	4	0.76
TESTING NONTRADITIONAL	10	4	0.91
TEXT EDITING CHARACTER	10	4	0.53
TEXT EDITING LINE	10	4	0.97
TEXT EDITING SCREEN	10	4	0.95

Scale Level Percentages of Agreement

Tables 6 and 7 show that at both the scale and subscale levels, mean agreement percentages are quite high, with only four of the Instruction subscales (Table 7): Planning (66%),

Student Documentation (66%), Interaction (70%) and Sequencing (70%) falling noticably below the 75% criterion. For total Scales, Physical (91%), Presentation (79%), Management (87%) and Instruction (80%) agreement percentages all fall above the criterion and indicate agreement among approximately nine of ten raters.

TABLE 6

Agreement Percentages Across Scores and Subscores

CITAR CCCEM Scores	Mean Percent of Agreement (4 raters, 10 CBTs)
TOTAL EFFICACY SCORE	**0.83**
Physical Score	**0.91**
Physical Screens	NA*
Physical Peripherals	0.91
Physical Concurrent Applications	NA
Presentation Score	**0.79**
Presentation Interface	NA
Presentation Text	0.73
Presentation Graphics	0.81
Presentation Sound	0.79
Presentaiton Design Features	0.85
Presentation Text Editing	0.77
Management Score	**0.87**
Management Documentation	0.77
Management Record Keeping	0.87
Management Data Entry/Analysis	0.92
Management Reports	0.87
Management Media	0.89
Management Access	NA
Management Security	0.78
Management Prescriptions	0.95
Management Browse/Sequencing	0.96
Instruction Score	**0.80**
Planning Subscore	0.66
Objectives Subscore	0.99
Techniques Subscore	0.80
Interaction Subscore	0.73
Resource Scope Subscore	0.84

* The items designated NA are determined from documentation by project manager prior to evaluation.

TABLE 7

Agreement Percentages Across Instruction Subscores

CITAR CCCEM Score	Mean Percent of Agreement (4 raters, 10 CBTs)	
Instruction Score	**0.80**	
Planning Subscore	0.66	
Instructive Student Documentation		0.66
Instructive Task Analysis		NA
Objectives Subscore	0.99	
Instructive Stated Goals		NA
Instructive Defined Outcomes		0.98
Instructive Defined Processes		1.00
Techniques Subscore	0.80	
Instructive Techniques		0.83
Instructive Presentation		0.74
Instructive Timing		0.87
Interaction Subscore	0.73	
Instructive Judging Wrong Response		0.74
Instructive Testing		0.82
Instructive Interaction		0.70
Instructive Questioning		0.75
Instructive Sequencing		0.70
Resource Scope Subscore	0.84	
Instructive Intrinsic Resource Scope		0.79
Instructive Extrinsic Resource Scope		0.90
Instructive Supplemental Materials		0.85

* The items designated NA are determined from documentation by project manager prior to evaluation.

RELIABILITY OF MAJOR SCORES

Regarding the ability of CCCEM scores to consistently discriminate across raters, Table 8 shows that the Efficacy score (.65, .77), the Management score (.67, .71) and the Instruction score (.66, .73) exhibit reasonable levels of reliability. Given the extremely conservative nature of these reliability estimates, there should be no reason to avoid the use of these scales for purposes of grouping CBT packages into four categories: **Outstanding, Excellent, Moderate** and **Questionable.** In order for a measure to differentiate consistently among CBT packages, it is necessary for those CBT packages to differ on the items of interest. Where useful differences are lacking, the instrument may exhibit high interrate agreement but cannot exhibit high reliability. The ranges of mean scores (mean of 4 raters) in Table 8 show that limited variability contributed to the lower reliability estimates for the Physical score (.55, .53) and the extremely low estimates obtained for the Presentation score (.17, .11). This also appears the major source of unreliability for the Planning subscore of instruction (.29, .47), although rater disagreement also contributed here (.66) and in the Interaction subscore (.73).

TABLE 8

Reliability of Major and Subscales

CITAR Score/Subscore	Range of Mean Scores	Reliability ICC	Reliability Tst/Rtst	Pct. Agrmt
Efficacy	34 - 57	.65	.77	.83
Physical Characteristics	66 - 80	.55	.53	.91
Presentation Aspects	38 - 56	.17	.11	.79
Management Structure	2 - 49	.67	.71	.87
Instruction Aspects	35 - 64	.66	.73	.80
Planning Subscore	18 - 30	.29	.47	.66
Techniques Subscore	35 - 81	.45	.54	.99
Objectives Subscore	13 - 83	.93	.94	.80
Interaction Subscore	14 - 59	.63	.63	.73
Resource Scope Subscore	48 - 95	.41	.53	.84

Another interesting aspect of the scores in Table 8, is the distance from optimum obtained by the highest scoring CBT packages for most scales. For four of the five major scores, no CBT package achieved two-thirds of the optimum score (.67). Only in the Physical score did any reach four-fifths of optimal (.80). Also, under Management, although considerable variability occurred, no CBT package even reached 50% of optimality. Among Instruction subscores, Planning and Interaction are particularly lacking among the CBT packages. It should be noted here that several subscores or items (e.g. task analysis) were found perhaps only once among the ten CBT packages included in this study. This despite the fact that the sample was chosen to include as many characteristics of the CCCEM system as possible.

ASSIGNMENT OF VALUES TO SCORES

Values (Good/Bad, High/Low) may be assigned to scores attained by CBT packages on the various subscales based upon either of two criteria: (1) Relative Performance or (2) Absolute Performance. Since each score represents a proportion of the optimal, either technique is applicable. For example, given an optimum score of 100, one might consider any CBT package attaining over 75% be defined as Outstanding. Or, given the relative scores of all CBT packages currently evaluated (n=180), one might consider the top 10/15 percent to be outstanding.

So few CBT packages currently approach optimum, it appears most appropriate to develop categories using relative percentiles:

1. **OUTSTANDING** - 85th to 99th percentile,
2. **EXCELLENT** - 60th to 84th percentile,
3. **MODERATE** - 40th to 59th percentile, and
4. **QUESTIONABLE**- 1st to 39th percentile.

Hopefully, over time, current CBT packages will be superceded by more sophisticated and efficacious software. Based upon the first 180 CBT packages, the following cutpoints appear appropriate (Appendix A contains source data for cutpoint development).

TABLE 9

Cut Scores for Rating CTBs

Score	Below Average	Average	Excellent	Outstanding
Efficacy	0 - 40	41 - 46	47 - 50	51 +
Instruction	0 - 47	48 - 54	55 - 63	64 +
Management		0 - 29	30 - 50	51 +
Presentation	Inadequate Reliability for Ratings			
Physical	0 - 65	66 - 71	72 - 77	78 +

COMPARISON OF RATER PERCEPTIONS AND CCCEM SCORES

In order to determine whether a gain in reliability (consistency) occurs by using relatively objective criteria such as those in the CCCEM model over pure evaluations by raters, the four evaluators also produced a quality rating for each CBT package corresponding to the four major CCCEM scores: **Instruction, Management, Presentation** and **Physical.** Ratings were based on a scale from zero to 10, where zero represents non-existent or useless and 10 represents optimal. Score based reliabilities using formula 1.0 were then computed for each of these and compared with similar results produced by the CCCEM scores. ICC estimates using formula 2.0 were also produced and compared between the two sources of information. As Table 10 shows, for agreement percentages, those produced by the CCCEM, even at the item level, average about

20 percentage points higher than those produced by pure rater perceptions. Perceptual disagreements among raters (shown in the ICC columns of Table 10), tend to invalidate an instrument. None of the perceptual scores showed the ability to consistently discriminate among CBT packages. With reliability estimates ranging from .08 to .23, it is not possible to determine which CBT package a particular rater is evaluating, since the within package variabilty is almost as the between package variability. Based upon this evidence, one may conclude that it is not possible to compare evaluations of two different software packages by two different evaluators based upon purely perceptual or opinion. The CCCEM, on the other hand, allows one to compare evaluations made by one rater with those of another on different software. Interestingly, Presentation was the least reliable score both perceputally and objectively.

TABLE 10

Consistency Estimates from Perceptions and CCCEM Scores

Score	Agreement		ICC	
	Perceptions	CCCEM	Perceptions	CCCEM
Efficacy	.69	.83	.18	.65
Instruction	.61	.80	.23	.66
Management	.64	.87	.10	.67
Presentation	.65	.79	.08	.17
Physcial	.69	.91	.11	.55

BIBLIOGRAPHY

Cook, T. D. and Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Chicago: Rand McNally

Frick, T. & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 48, 157-184.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and the generalizability of data collected in observation studies. Psychological Bulletin, 86, 376-390.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. Psychological Bulletin, 81, 420-428.

Towstopiat, O. (1984). A review of reliability procedures for measuring observer agreement. Contemporary Educational Psychology, 9, 333-352.

Appendix A

DISTIBUTION OF CITAR SCORES ATTAINED BY CURRENT CBT PACKAGES

The following tables show the scores obtained by all CBT packages evaluated by CITAR as of August, 1987 on a scale from zero to 100. Table 10 indicates that although Efficacy scores exhibit variability (s_x = 8.416), the mean (43.36) and range (20 - 64) of scores shows considerable room for improvement. Physical scores (Table 11) exhibit limited variability (s_x = 6.42) and show relatively higher proportions of optimality than other scales (mean = 71.11). Presentation scores (Table 12) range substantially (20 to 65). Management scores (Table 13) suggest that few CBT packages exhibit many management system characteristics, with 84% receiving less than 10% of the possible score. Of those having management systems, fewer than 10% exhibit as much as 40% of the optimum capicity. Table 14 shows the overall instruction score, which ranges from 17 to 77, has a mean of 51.18 and exhibits adequate variability (s_x = 11.99).

TABLE 11

Total Efficacy Score - Frequencies for 180 CBTs

Score	Frequency	Percent	Cumulative Percent
20	1	.6	.6
23	1	.6	1.1
25	1	.6	1.7
27	1	.6	2.2
28	3	1.7	3.9
29	1	.6	4.4
30	1	.6	5.0
31	6	3.3	8.3
32	3	1.7	10.0
33	6	3.3	13.3
34	6	3.3	16.7
35	2	1.1	17.8
36	9	5.0	22.8
37	9	5.0	27.8
38	4	2.2	30.0
39	6	3.3	33.3
40	6	3.3	36.7
41	11	6.1	42.8
42	4	2.2	45.0
43	5	2.8	47.8
44	8	4.4	52.2
45	7	3.9	56.1
46	5	2.8	58.9
47	14	7.8	66.7
48	10	5.6	72.2
49	11	6.1	78.3
50	7	3.9	82.2
51	6	3.3	85.6
52	2	1.1	86.7
53	4	2.2	88.9
54	5	2.8	91.7
55	3	1.7	93.3
56	2	1.1	94.4
57	3	1.7	96.1
59	2	1.1	97.2
60	1	.6	97.8
62	1	.6	98.3
63	1	.6	98.9
64	2	1.1	100.0
TOTAL	180	100.0	

MEAN	43.356	MEDIAN	44.000
SKEWNESS	-.038	RANGE	44.000
STD DEV	8.416		

TABLE 12

Total Physical Score - Frequencies for 180 CBTs

Score	Frequency	Percent	Cumulative Percent
52	1	.6	.6
57	8	4.4	5.0
63	2	1.1	6.1
64	5	2.8	8.9
66	49	27.2	36.1
69	5	2.8	38.9
70	5	2.8	41.7
72	53	29.4	71.1
73	2	1.1	72.2
75	2	1.1	73.3
76	6	3.3	76.7
78	30	16.7	93.3
82	8	4.4	97.8
88	4	2.2	100.0
TOTAL	180	100.0	

MEAN	71.106	MEDIAN	72.000
SKEWNESS	.031	RANGE	36.000
STD DEV	6.424		

TABLE 13

Total Presentation Score - Frequencies for 180 CBTs

Score	Frequency	Percent	Cumulative Percent
20	1	.6	.6
23	1	.6	1.1
27	1	.6	1.7
28	2	1.1	2.8
29	1	.6	3.3
32	4	2.2	5.6
33	3	1.7	7.2
34	3	1.7	8.9
35	5	2.8	11.7
36	5	2.8	14.4
37	9	5.0	19.4
38	8	4.4	23.9
40	3	1.7	25.6
41	9	5.0	30.6
42	7	3.9	34.4
43	15	8.3	42.8
44	7	3.9	46.7
45	11	6.1	52.8
46	8	4.4	57.2
47	3	1.7	58.9
48	5	2.8	61.7
49	7	3.9	65.6
50	8	4.4	70.0
51	6	3.3	73.3
52	14	7.8	81.1
53	3	1.7	82.8
54	8	4.4	87.2
55	11	6.1	93.3
56	2	1.1	94.4
57	2	1.1	95.6
58	1	.6	96.1
59	2	1.1	97.2
60	1	.6	97.8
64	1	.6	98.3
65	3	1.7	100.0
TOTAL	180	100.0	

MEAN	45.278	MEDIAN	45.000
SKEWNESS	-.176	RANGE	45.000
STD DEV	8.202		

TABLE 14

Total Management Score - Frequencies for 180 CBTs

Score	Frequency	Percent	Cumulative Percent
1	51	28.3	28.3
2	73	40.6	68.9
3	15	8.3	77.2
4	10	5.6	82.8
6	1	.6	83.3
7	1	.6	83.9
21	1	.6	84.4
26	1	.6	85.0
28	1	.6	85.6
34	1	.6	86.1
36	1	.6	86.7
37	2	1.1	87.8
38	3	1.7	89.4
39	3	1.7	91.1
41	1	.6	91.7
43	2	1.1	92.8
44	4	2.2	95.0
45	2	1.1	96.1
46	2	1.1	97.2
47	1	.6	97.8
48	2	1.1	98.9
50	1	.6	99.4
54	1	.6	100.0
TOTAL	180	100.0	

MEAN	8.206	MEDIAN	2.000
SKEWNESS	1.979	RANGE	53.000
STD DEV	14.614		

TABLE 15

Total Instruction Score - Frequencies for 180 CBTs

Score	Freq	Percent	Cumulative Percent	Score	Freq	Percent	Cumulative Percent
17	1	.6	.6	59	8	4.4	73.9
22	1	.6	1.1	60	4	2.2	76.1
23	1	.6	1.7	61	4	2.2	78.3
25	1	.6	2.2	62	3	1.7	80.0
29	3	1.7	3.9	63	7	3.9	83.9
31	2	1.1	5.0	64	4	2.2	86.1
32	2	1.1	6.1	65	1	.6	86.7
33	2	1.1	7.2	66	6	3.3	90.0
34	5	2.8	10.0	67	3	1.7	91.7
35	3	1.7	11.7	68	3	1.7	93.3
36	4	2.2	13.9	69	3	1.7	95.0
37	2	1.1	15.0	70	4	2.2	97.2
38	3	1.7	16.7	71	1	.6	97.8
39	4	2.2	18.9	73	1	.6	98.3
40	2	1.1	20.0	74	1	.6	98.9
41	1	.6	20.6	75	1	.6	99.4
42	2	1.1	21.7	77	1	.6	100.0
43	8	4.4	26.1				
44	8	4.4	30.6	TOTAL	180	100.0	
45	3	1.7	32.2				
46	3	1.7	33.9				
47	8	4.4	38.3				
48	5	2.8	41.1				
49	4	2.2	43.3				
50	7	3.9	47.2				
51	7	3.9	51.1				
53	5	2.8	53.9				
54	9	5.0	58.9				
55	2	1.1	60.0				
56	6	3.3	63.3				
57	6	3.3	66.7				
58	5	2.8	69.4				

MEAN	51.183	MEDIAN	51.000	STD DEV	11.991
SKEWNESS	-.251	RANGE	60.000		

Appendix B

ITEM CONTENT OF SCALES AND SUBSCALES

PHYSICAL SUBSCALES

PHYSICAL SCREENS

PA01	HARD DISKS
PA02	RESOLUTION GIVEN
PA03	BOARDS
PA04	TWO SCREENS
PA05	COLOR
PA06	80 CHARACTERS
PA07	40 CHARACTERS
PA08	UPPER AND LOWER CASE

PHYSICAL PERIPHERALS

PB01	PERIPHERALS
PB02	PERIPHERALS PRINTER
PB03	PERIPHERALS PLOTTER
PB04	PERIPHERALS EXT HOST
PB05	PERIPHERALS AUDIOTAPE
PB06	PERIPHERALS VIDEODISC
PB07	PERIPHERALS SLIDE
PB08	PERIPHERALS VOICE RECOGNITION
PB09	PERIPHERALS CD-ROM
PB10	PERIPHERALS GAME ADAPTOR
PB11	PERIPHERALS MODEM

Presentation Interface

PJ01	INTERFACE MOUSE
PJ02	INTERFACE TOUCH
PJ03	INTERFACE KEYBOARD
PJ04	INTERFACE LIGHT PEN
PJ05	INTERFACE BIT/GRAPHICS PAD
PJ06	INTERFACE MICROPHONE

PRESENTATION SUBSCORES

Presentation Text

PD01	TEXT COLORS
PD02	TEXT SIZING
PD03	TEXT FONTS
PD04	TEXT FORMATS
PD05	TEXT FLASH OR BLINK
PD06	TEXT PRINTOUTS
PD07	TEXT SCREEN DUMP

Presentation Graphics

PE01	GRAPHICS COLOR
PE02	GRAPHICS ANIMATION
PE03	GRAPHICS VIDEO
PE04	GRAPHICS WINDOWING
PE05	GRAPHICS FLASH OR BLINK
PE06	GRAPHICS PRINTOUT
PE07	GRAPHICS SCREEN DUMP

PRESENTATION SOUNDS

PF01	SOUNDS COMPUTER TONES
PF02	SOUNDS EXTERNAL SOURCES
PF03	SOUNDS MUSIC
PF04	SOUNDS SYNTHESIZED VOICE

Presentation Design Features

PG01	DESIGN FEATURES SCREEN LAYOUT
PG02	DESIGN FEATURES LESSON FLOW
PG03	DESIGN FEATURES COLOR
PG04	DESIGN FEATURES SOUND

Presentation Text Editing

PH01	TEXT EDITING CHARACTER
PH02	TEXT EDITING LINE
PH03	TEXT EDITING SCREEN

MANAGEMENT SUBSCORES

Management Documentation

MA01	DOCUMENTATION MANAGER
MA02	DOCUMENTATION MANUALS
MA03	DOCUMENTATION TABLES OF CONTENTS
MA04	DOCUMENTATION INDEX
MA05	DOCUMENTATION INSTRUCTOR
MA06	DOCUMENTATION MANUALS
MA07	DOCUMENTATION TABLES OF CONTENTS
MA08	DOCUMENTATION INDEX

Management Record Keeping

MB01	MS RECORD KEEPING?
MB05	TESTING HISTORY
MB06	UNANTICIPATED RESPONSES
MB07	COURSEWARE VALIDITY

Management Data Analysis

MC01	MS DATA ENTRY AUTOMATIC
MC02	MS DATA ENTRY MANUAL
MC03	MS DATA ANALYSIS
MC04	MS DATA ANALYSIS INTERNAL
MC05	MS DATA ANALYSIS TEST
MC06	MS DATA ANALYSIS ITEM
MC07	MS DATA ANALYSIS EXTERNAL

Management Reports

MD01	MS REPORTS
MD02	MS REPORTS PRE FORMATTED
MD03	MS REPORTS USER FORMATTED

Management Media

ME01	MS MEDIA PAPER BASED
ME02	MS MEDIA FLOPPY
ME03	MS MEDIA NETWORK
ME04	MS MEDIA MAINFRAME

Management System Access

ME05	MS ACCCEMSS
ME06	MS ACCCEMSS TERMINAL
ME07	MS ACCCEMSS MICRO
ME08	MS ACCCEMSS HARDCOPY
ME09	MS ACCCEMSS MANAGER ONLY

Management System Security

MF01	MS SECURITY
MF02	MS SECURITY STUDENT ACCCEMSS?

Management Learning Prescriptions

MG01	MS LEARNING PRESCRIPTIONS
MG02	MS LEARNING PRESCRIPTIONS FORMATIVE
MG03	MS LEARNING PRESCRIPTIONS COURSE LEV
MG04	MS LEARNING PRESCRIPTIONS MODULE LEV
MG05	MS LEARNING PRESCRIPTIONS ADAPTATION

Management Instructor Browse

MH01	INSTRUCTOR BROWSE
MH02	INSTRUCTOR SEQUENCING
MH03	INSTRUCTOR GROUP
MH04	INSTRUCTOR INDIVIDUAL
MH05	OPTIONAL SEQUENCE TO HELP
MH06	OPTIONAL SEQUENCE TO HINT
MH07	OPTIONAL SEQUENCE TO TUTORIAL
MH08	OPTIONAL SEQUENCE TO TEST
MH09	OPTIONAL SEQUENCE TO MENU
MH10	OPTIONAL SEQUENCE TO OFF LINE
MH11	TESTING VARIABLE SCORING OPTIONS
MH12	TESTING UNOBTRUSIVE

INSTRUCTION SUBSCORES

Instruction Student Documentation

IA01	DOCUMENTATION STUDENT
IA02	DOCUMENTATION MANUALS
IA03	DOCUMENTATION TABLES OF CONTENTS
IA04	DOCUMENTATION INDEX

Instruction Defined Goals

IC01	OBJECTIVES ENTRY PROFICIENCY
IC02	OBJECTIVES EXIT PROFICIENCY

Instruction Defined Outcomes

ID01	OUTCOME DEFS COGNITIVE
ID02	OUTCOME DEFS MEASURED
ID03	OUTCOME DEFS AFFECTIVE
ID04	OUTCOME DEFS MEASURED
ID05	OUTCOME DEFS PSYCHOMOTOR
ID06	OUTCOME DEFS MEASURED

Instruction Defined Processes

IE01	PROCESS DEFS COGNITIVE
IE02	PROCESS DEFS DOCUMENTED
IE03	PROCESS DEFS AFFECTIVE
IE04	PROCESS DEFS DOCUMENTED
IE05	PROCESS DEFS PSYCHOMOTOR
IE06	PROCESS DEFS DOCUMENTED

Instruciton Techniques

II01	PAGE TURNING
II02	TUTORIAL
II03	DRILL AND PRACTICE
II04	QUESTIONING
II05	CONCURRENT TRAINING
II06	PROBLEM SOLVING
II07	SIMULATION
II08	MODELING
II09	GAMING
II10	COOPERATIVE GAMING
II11	INQUIRY

Instruction Physical Presentation

IJ01	GRAPHS, CHARTS ETC
IJ02	ILLUSTRATIONS

Instruction Timing

IK01	TIMING FORCED WAIT
IK02	OVERRIDE
IK03	FORCED MOVEMENT
IK04	OVERRIDE
IK05	DISPLAY INTERRUPT

Instruction Judging Incorrect Responses

IN01	WAIT FOR CORRECT RESPONSE
IN02	WRONG AND CONTINUE
IN03	WRONG AND HINT
IN04	WRONG AND HELP
IN05	WRONG AND TUTORIAL
IN06	WRONG AND OFF LINE
IN07	SCORE AND GO
IN08	GIVE CORRECT AND GO
IN09	STUDENT TYPES CORRECT RESPONSE
IN10	ANSWER KEY PROVIDED

Instruction Testing

IP01	TESTING RANDOM
IP02	TESTING FIXED
IP03	TESTING NONTRADITIONAL

Instruction Interaction

IL01	INTERACTION PROMPTS
IL02	INTERACTION FEEDBACK
IL03	INTERACTION FEEDBACK POSITIVE
IL04	INTERACTION FEEDBACK NEGATIVE
IL05	INTERACTION FEEDBACK NEUTRAL
IL06	INPUT QUITTING
IL07	INPUT RESTART
IL08	RESPONSE JUDGING
IL09	ONLY ONE RESPONSE
IL10	SLIGHT ERRORS
IL11	SENTENCE PARSING
IL12	LINGUISTIC
IL13	MATHEMATICAL
IL14	NATURAL LANGUAGE

Instruction Questioning

IM01	QUESTIONS?
IM02	QUESTIONS MULTIPLE CHOICE
IM03	QUESTIONS MATCHING
IM04	QUESTIONS TRUE FALSE
IM05	QUESTIONS FILL BLANK
IM06	QUESTIONS NUMERIC RESPONSE
IM07	QUESTIONS GRAPHIC
IM08	QUESTIONS FREEFORM TEXT
IM09	QUESTIONS CONTINUATION RESPONSE
IM10	QUESTIONS ON LINE HELP
IM11	QUESTIONS CONTEXT SENSITIVE

Instruction Sequencing

IO01	SEUQENCE LINEAR
IO02	SEUQENCE DUMMY KEY
IO03	SEUQENCE BRANCHING
IO04	SEUQENCE DIRECTION
IO05	SEUQENCE BACKWARD?
IO06	SEUQENCE ONE PAGE BACKWARD
IO07	SEUQENCE MULTIPLE PAGES BACKWARD
IO08	SEUQENCE FOWARD
IO09	SEUQENCE ONE PAGE FORWARD
IO10	
IO11	SEUQENCE MASTERY
IO12	SEUQENCE CREDIT FOR MASTERY
IO13	SEUQENCE ADAPTIVE
IO14	SEUQENCE BROWSE

Instruction Resource Scope Intrinsic

IF01	RESOURCE SCOPE INTRINSIC
IF02	TEXT
IF03	GRAPHICS
IF04	ANIMATION
IF05	SOUND

Instruction Resource Scope Extrinsic

IG01	RESOURCE SCOPE EXTRINSIC
IG02	BOOKS
IG03	LABS
IG04	SEMINARS
IG05	LECTURES
IG06	FIELD TRIPS
IG07	HELP SESSIONS
IG08	TECHNOLOGICAL

Instruction Supplemental

IH01	SUPPLEMENTAL
IH02	BOOKS
IH03	LABS
IH04	SEMINARS
IH05	LECTURES
IH06	FIELD TRIPS
IH07	HELP SESSIONS
IH08	TECHNOLOGICAL